Proceedings of the Institute of Acoustics

PHYSIOLOGICALLY CONTROLLED VOICE SOURCE MODELS FOR DIFFERENT SPEAKERS

Celia Scully, Karen Stromberg

University of Leeds, Department of Psychology, Leeds, England

1. TERMINAL ANALOG SYNTHESIS AND ARTICULATORY SYNTHESIS

Terminal-analog synthesis is based upon the approximation that acoustic sources and filters are all independent of each other; in reality they are interdependent because of the patterns of air movement shared by all the various acoustic sources and by the vocal tract filter.

For the purposes of this paper two different types of air movement and air pressure will be distinguished: low frequencies from dc up to about 50 Hz for the aerodynamic processes of speech production and audio frequencies from about 50 Hz upwards to describe acoustic sources, filtering and soundwave radiation. The formant frequencies and bandwidths needed for terminal-analog synthesis are on the whole easy to obtain: they can be copied from real speech individually. Care must be taken to avoid unnatural combinations of formant frequencies. Naturalness of formant patterning is improved automatically in line-analog synthesis because of the constraint that all the formants arise from simultaneous resonances of a single vocal tract tube.

It seems logical to pursue this concept further and to assert that, as more and more of the constraints of real speech are built into a synthesiser, the naturalness of its output increases. This is the justification for considering articulatory synthesis as a potentially useful approach to artificial voices. If the configurations and movement paths of the articulatory stage of speech production can be described and if the physical processes which generate the acoustic sources and which filter them can be represented in a simplified way, then a great deal of the richness and complexity of real speech will be built into the resulting synthesis, a better signal model will be employed and the synthetic speech will sound more natural.

The difficulties lying in the path of the two requirements just stated are considerable and may well preclude the use of articulatory synthesis as a marketable method. But searches for good matches to articulation and to the physical mechanisms of speech production can themselves aid our understanding of the acoustic structures of speech signals. A more complete knowledge of speech production processes can provide rules for terminal-analog synthesis to use, thus improving the naturalness of this kind of synthetic speech.

We focus on the voice source primarily, but sources located at the larynx should not be considered in isolation. The three types of acoustic source to be considered are voice, noise - aspiration noise generated just above the glottis and frication noise generated in front of the outlet of extremely constricted regions of the vocal tract, and transient generated when two regions at different air pressures are brought into contact.

2.1. VOICE AND OTHER SOURCES: ACOUSTIC INTERACTIONS BETWEEN SOURCE AND FILTER

Better naturalness seems to be obtained for synthetic voices when they include the acoustic loading of the resonances of the vocal tract filter upon the oscillatory action of the vocal folds, and thus upon the voice source (Pinto and Childers [1]). In vocoid (vowel-like) sounds different vocal tract area functions contribute to different amounts of skewing of the voice source waveshape (Rothenberg [2]). In addition, the larynx actions associated with the properties of the various acoustic sources have an effect on the acoustic filter. The glottis must be noticeably enlarged for the production of successful voiceless fricatives and the voiceless aspirated plosives of a language such as English; it needs to be slightly enlarged for voiced fricatives and probably for other kinds of plosives also. In this case the acoustic losses across the glottis increase, and so the formant bandwidths increase. The effects should be apparent in vocoid-contoid boundary regions, as for a vowel followed by a voiceless fricative. Here the voice source is becoming weaker and losing its high frequencies while noise sources are increasing in strength. In this region voice, aspiration noise, frication noise and sometimes a transient also occur almost simultaneously; in their overlap portions the unity of the acoustic filter is expressed through the formant transitions.

2.2. VOICE AND OTHER SOURCES : AERODYNAMIC INTERACTIONS BETWEEN SOURCE AND FILTER

All portions of the respiratory tract, from the lungs to the mouth and nose outlets, are linked by the air flowing through them. Aerodynamic forces which, together with mechanical forces such as tissue elasticity, cause the vocal folds to oscillate are associated with this unified, irreducible aerodynamic circuit. It behaves very like an electric circuit, except that in the place of Ohm's law $V = IR$ is the orifice equation applied to sections of severe constriction of the respiratory tract, including the glottis, at which there is significant resistance to airflow:

$$A = K.U/\surd(\Delta P) \qquad (1)$$

where A is the cross-section area of the constriction, U is the volume flowrate of air through it, ΔP is the air pressure drop across it and K is an empirical constant (Scully [3]): U and ΔP are aerodynamic components.

In this aerodynamic circuit, the airflow UG through the glottis is closely related to, though not always identical to, the airflow UC through a constriction of the vocal tract, for example a labio-dental constriction as in the production of [f] and [v]. For a non-nasal configuration of the vocal tract the aerodynamic circuit can be represented as in Figure 1.

AG is the average articulatory, non-oscillatory component of glottal area. When the vocal tract constriction AC is made more severe, so that AC decreases, oral air pressure in the vocal tract PC rises. The pressure drop across the glottis (PSG - PC) is therefore reduced (assuming that subglottal pressure remains more or less constant) and with it the aerodynamic forces which make the vocal folds oscillate, so that voicing becomes acoustically weaker and, if PC rises to a value nearly as high as PSG, ceases altogether.

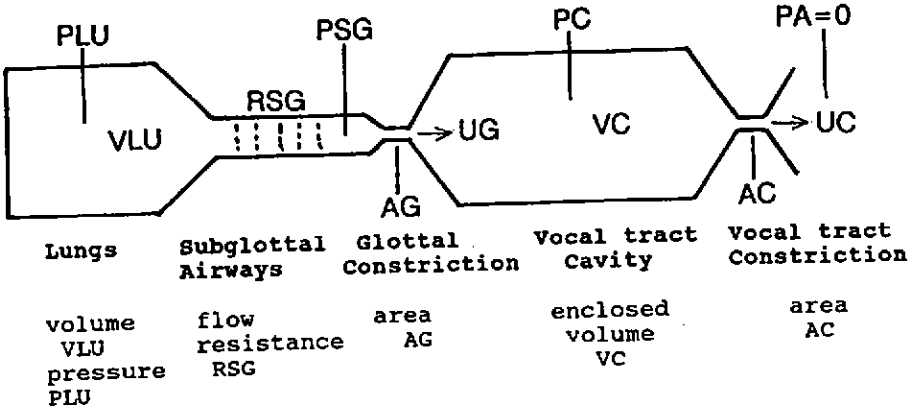


Figure 1. Schematic representation of a non-nasal aerodynamic circuit.

Considering the larynx-vocal tract interactions in the other direction: if the glottal area is enlarged, the flow of air UG into the vocal tract enclosed volume VC increases. If at the same time the cavity outlet constriction forms a severe obstruction with small AC, then air pressure PC inside the enclosed volume rises rapidly, more or less rapidly depending upon how much the glottis is enlarged. The rate of rise in oral air pressure PC is thus determined by both the larynx and the vocal tract actions. In a model of speech production these aerodynamic effects can be represented quantitively by the solution of a set of simultaneous differential equations, progressing along the time base (Scully & Allwood [4]). If the glottis is enlarged sufficiently then voicing becomes weaker acoustically and ceases, even though the vocal tract is not constricted. The rate of oscillation of the vocal folds depends mainly upon their effective vibrating mass and stiffness combined; this parameter is labelled Q in models of voicing, and in this paper also. There is an aerodynamic factor also (Ladefoged [5]).

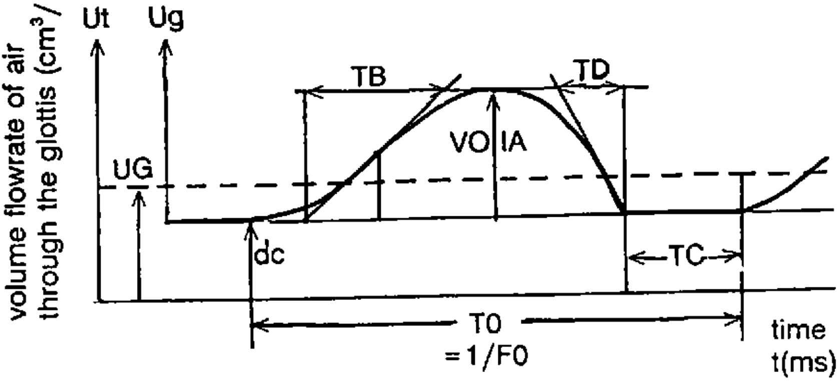
If a single aerodynamic parameter were to be picked out as linking the voice source, aspiration noise, frication noise and transient with the area function of the vocal tract filter, this could be PC, the air pressure inside the vocal tract cavity. Its value, combined with AC, determines the strength and frequency spectrum of frication noise; its rate of change is probably the main controlling factor for the transient source; it interacts with the subglottal air pressure PSG to affect the voicing mechanism and to determine the strength of the aspiration noise source; its value and overall pattern as a function of time is a reflection of both vocal tract and larynx articulations, the former being the main determinant of formant frequencies and the latter being one of the controlling factors for the generation of the voice source and with an influence on formant bandwidths.

3. MODELLING THE VOICE SOURCE : A PHENOMENOLOGICAL APPROACH

We have attempted to demonstrate that a few aerodynamic parameters, arising from and interacting with articulatory states and movement, control many acoustic pattern features of the speech signal, for both vowels and consonants. All these details of signal structure, and in particular the fact that they covary because of the common physiological controlling conditions which they share, seem likely to provide useful information for listeners, to help them decode the linguistic message and receive speaker-specific information also. We focus here on one aspect only, the way the voice source changes in the time domain with phonetic context. We attempt to characterise the range of waveshapes for two different speakers, thus placing individual-speaker parameter values within a standardised descriptive framework. Two speakers were analysed: CS, a woman speaker of General American, and PB, a man speaker of French from Grenoble.

Some of the best-known models of voicing are true physical models of the processes, for example, the two-mass model of Ishizaka and Flanagan [6]. Another approach is to characterise the voice source waveshape, that is the acoustic component of airflow through the glottis, under different conditions and for different speakers (for example Fant [7]). We have adopted a phenomenological approach for the voice source in our articulatory synthesis based upon the voice waveshape parameters shown in Figure 2. This model does not allow for a rounded corner at closure as is done in the L-F model (Fant et al. [8]). The parameters needed to construct the voice source waveshape are the following :

VOIF (Hz)	fundamental frequency = F0 (=1/T0)
VOIA (cm^3/s)	amplitude of wave
TCR	duration of closed phase TC/total cycle duration T0
K	asymmetry factor defined as shown in Figure 2
TD (ms)	duration of closing portion as shown in Figure 2



$K = 0.5 + 0.125(TB/TD)^2$ VOIF = F0 TCR = TC/T0

Figure 2 Parametric representation of the voice source waveform (based on Fant [7]).

VOIF and three of the other four are selected to define the voice waveshape. Each of these is assumed to be determined by three controlling physiological parameters :

PDIFF (cmH$_2$0) the pressure drop across the glottis (PSG-PC)
Q (Hz) the effective mass and stiffness of the vocal folds
AG (cm^2) the articulatory (non-oscillatory) component of glottal area

The option VOIA, TCR, K is described here although TD has been included elsewhere in our modelling of singing (Scully & Allwood [9]). The mapping from physiological parameters to fundamental frequency of the voice waveshape is taken to be the following :

$$\text{VOIF} = \text{Q} + \text{KF.PDIFF} \qquad (2)$$

KF is an empirical constant to be determined for each speaker. Published studies suggest a value near 4 for KF (Ladefoged [5]) but quite a wide range of values can be found in the literature.

Dimension	value	file name	physiologic.parameter expected	voice waveshape expected
effort level	medium	lm	med PDIFF	ordinary
	soft	ls	low PDIFF	weak
	loud	ll	high PDIFF	strong
pitch	mid	fm	mid Q	mid (normal) F_0
	low	fl	low Q	low F_0
	high	fh	high Q	high F_0
phonation	normal	pn	normal AG	normal
	pressed	pp	small AG	laryngeal
	breathy	pb	large AG	weak, breathy

Table I The nine different voicing types used to establish the range of voice waveshapes under different combinations of physiological controlling variables. The descriptions in the table are labels for what may be complex effects. lm, fm and pn should all give similar voice waveshapes since all three types have mid values for level, pitch and phonation.

We obtain estimates for the voice source waveshape Ug(t) under several different sets of physiological conditions. We try to elicit from the real speakers productions that are widely distributed across the three-dimensional space for PDIFF, Q and AG. To this end, the speaker produces nine series of [pV pV pV ...] on a single expiratory breath as shown in Table I. [V] should be a vocoid which does not require a strongly lowered jaw. The reason is that jaw movements between the high jaw needed for [p] and a low-jaw vocoid increase the difference between the total output volume flowrate of air (Uo + Un) and the total volume flowrate of air through the glottis, Ut (see Figure 2). Uo, the oral component, and Un, the nasal component, are added to give the total output flow, Uto, with aerodynamic and acoustic components both included.

4. DATA ACQUISITION AND ANALYSIS

Uto was obtained using an undivided Rothenberg airflow mask. Oral air pressure PC was obtained using an orally-inserted pressure tube with a Gaeltec pressure transducer. The speech signal was obtained with a B&K microphone outside the mask and a laryngograph signal was obtained at the same time. The signals were recorded on digital video cassette (Sony) and on 4-track FM recorder (Racal).

The four channels of data were logged to a Masscomp 5500 computer from the FM recorder. Calibration signals for airflow and air pressure were logged also. The speech corpus and the methods are described in detail elsewhere (Guérin et al [10]). The output airflow signal was inverse filtered (Brookes et al [11]) to obtain estimates for Ut(t) and Ug(t). Generally, three repetitions of each voicing type shown in Table I were analysed, excluding the first in the series, at a near mid-vocoid time point. Average glottal airflow UG and average pressure drop across the glottis PDIFF (PSG - PC) were estimated from mingograph (hard copy) traces. These aerodynamic parameters could be related to acoustic parameters, especially F0, and also to the voice waveshape.

Since total airflow was used as the input, the inverse filtering gave estimates for both acoustic and dc components of glottal flow as shown in Figure 2. UG was calculated by assuming a triangular shape for the open phase. This often seemed a good shape approximation as seen in Figure 3 which shows one waveshape for speaker CS. The equation used was :

$$UG = dc + 0.5\,(1-TCR)\,VOIA \qquad (3)$$

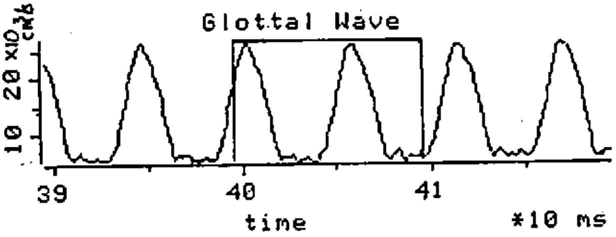


Figure 3 An example of the voice source waveshape obtained by inverse filtering : low pitch, speaker CS.

This was compared with UG as measured directly on the mingogram. The consistency seemed reasonably acceptable, given the difficulties of estimating parameter values for the voice waveshape and our non-specialist use of inverse filtering as a technique. As good and bad examples : for speaker CS lm2 gave 100 and 104, fl4 gave 50 and 27; for speaker PB pn4 gave 115 and 118, ft3 gave 60 and 105 cm^3/s.

5. CONSTRUCTION OF THE LINEAR MAPPING FROM PHYSIOLOGICAL CONDITIONS TO VOICE SOURCE WAVESHAPE PARAMETERS

PSG was estimated near each mid-vocoid point by linear interpolation between peak oral air pressure values for the [p] plosives on either side (Scully [3]).

PDIFF was obtained by subtracting PC at that time point. PC was not noticeably different from zero (atmospheric pressure), except in the case of breathy phonation. PDIFF and UG (taken as equal to total output flow UC - see Figure 1) were combined in the orifice equation, equation (1) above, to give an estimate for AG.

Q and the empirical constant KF were obtained from the dip in fundamental frequency associated with the speaker's production of voicing during a voiced fricative under reduced transglottal pressure PDIFF. Several measurements of F0 (VOIF), with an estimate for PDIFF were made over a short time span, both into and out of several voiced fricatives in a series [pVFV:pVFV:---] where F is a fricative, and a graph of F0 versus PDIFF was plotted. Assuming the linear relationship of equation (2) above, KF (units Hz/cmH_2O)is the gradient and Q is the intercept on the ordinate. A different value for KF for the two speakers was indicated, 8 for CS and 1 for PB. Their Q values differed also as would be expected for a woman and a man speaker.

Scattergrams were obtained for each of the voice waveshape parameters VOIA, TCR or K plotted against one controlling parameter PDIFF, Q or AG. Two examples are shown in Figure 4 for PB.

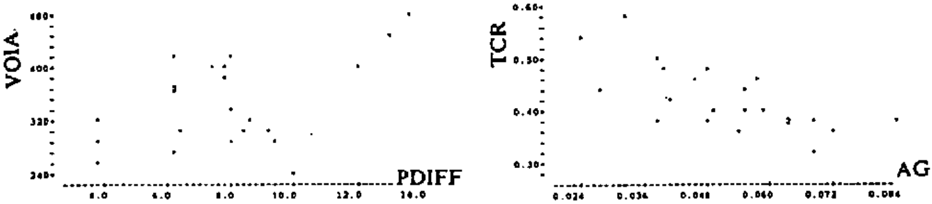


Figure 4 Two scattergrams for speaker PB.

Multivariate regression (Minitab package) was used for each waveshape parameter in turn. The equations obtained were :

For speaker CS:

$\mathrm{VOIA} = 4.85\ \mathrm{PDIFF} + 225\ \mathrm{AG} - 1.88\ Q + 435$
$\mathrm{TCR} = 0.0305\ \mathrm{PDIFF} - 2.64\ \mathrm{AG} + 0.00098 + 0.129$
$K = 0.0629\ \mathrm{PDIFF} + 4.32\ \mathrm{AG} + 0.00961\ Q - 1.26$

For speaker PB:

$\mathrm{VOIA} = 11.4\ \mathrm{PDIFF} - 330\ \mathrm{AG} - 0.13\ Q + 285$
$\mathrm{TCR} = 0.00427\ \mathrm{PDIFF} - 2.66\ \mathrm{AG} - 0.000163\ Q + 0.541$
$K = -0.0465\ \mathrm{PDIFF} - 0.7\ \mathrm{AG} + 0.0107\ Q + 0.22$

Other parameters are needed for the complete voicing model and the generation of noise sources. They are described in Guérin et al. [10]). With the complete voicing model incorporated into our composite model of speech production processes, simulations of a particular speaker produce automatic onset and cessation of voice and other sources at physiologically realistic time points, as well as a voice source spectrum which varies with phonetic context.

We are trying to improve the voice model by using stepwise multivariate regression. Stepwise regression finds the line of best fit (expressed as a multivariate regression equation) by discarding the most non-significant variable from the analysis each time and repeating the regression analysis until only the significant variables (the optimal set) remain in the equation.

The three multivariate regression equations thus obtained for speaker PB are given below :-

$\mathrm{VOIA} = 247 + 12.3\ \mathrm{PDIFF}$
$\mathrm{TCR} = 0.573 - 3.00\ \mathrm{AG}$
$K = 0.151 - 0.0446\ \mathrm{PDIFF} + 0.0109\ Q$

For PB at least the mappings from physiological controls to VOIA and TCR appear to be simple ones. We need to simulate both speakers with our articulatory synthesis in order to tune the parameter values, some of which are derived from rough estimates based on incomplete information from the real speech. We hope that these and other speakers can be characterised in a way which will help to generate synthetic speech having appropriate complexity and covariation of acoustic pattern features and with individual speaker characteristics also.

6. ACKNOWLEDGMENTS

Work done as part of an EC project SCI 0147-C (EDB). We appreciate this collaboration with colleagues at ICP Grenoble and at the Department of Electronics and Computer Science, Southampton. We thank also Gustav Clark and Robin Haigh who developed the software for data logging and calibration, Eric Castelli and Esther Grabe who worked on the project at Leeds, also Mike Brookes and other colleagues in the Alvey SPAR project under which the inverse filtering was developed, and Nawal Ghali for her advice and help with multivariate regression.

7. REFERENCES

[1] N B PINTO & D G CHILDERS, 'Formant Speech Synthesis', J Inst Electr and Telecom Engs, 34 pp 5-20 (1988)

[2] M ROTHENBERG, 'Acoustic Interaction Between the Glottal Source and the Vocal Tract', in 'Vocal Fold Physiology', STEVENS & HIRANDO eds, Univ of Tokyo Press pp 305-323 (1981)

[3] C SCULLY, 'Speech Production Simulated with a Functional Model of the Larynx and the Vocal Tract', J Phonetics 14 pp 407-414 (1986)

[4] C SCULLY & E ALLWOOD, 'Lung and Larynx Coordination in a Composite Model of Speech Production' Proc Tenth Intl Congr Phon Sci, VAN DEN BROECKE & COHEN eds, Foris Dordrecht pp 372-377 (1984)

[5] P LADEFOGED, 'Some Physiological Parameters in Speech' Lang & Speech 6 pp 109-119 (1963)

[6] K ISHIZAKA & J L FLANAGAN, 'Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords' Bell Syst Tech J 51 pp 1233-1268 (1972)

[7] G FANT, 'Voice Source Dynamics', STL-QPSR Stockholm 2-3/1980 pp17-37 (1980)

[8] G FANT, J LILJENCRANTS & Q LIN, 'A Four-Parameter Model of Glottal Flow', STL_QPSR Stockhlom 4/1985 pp 1-13 (1985)

[9] C SCULLY & E ALLWOOD 'Simulation of Singing with a Composite Model of Speech Production', A ASKENFELT et al. eds Royal Swedish Academy of Music Stockholm Vol I pp 247-259 (1985)

[10] B GUERIN et al., 'Mesure, Caractérisation et Modélisation des Sons Fricatifs', Final Report for Contract SCI*0147-C (EDB), ICP, Grenoble (1992)

[11] D M BROOKES, D M HOWARD & D S F CHAN, 'Dynamic Excitation Control in Parallel Formant Speech Synthesis', FASE88, Edinburgh Vol3 pp 1123-1130 (1989)